Communication in restricted situations: the use of orofacial expressions in different speech modes and when wearing a face mask

Nasim Mahdinazhad Sardhaei[1], *Marzena Zygis* [1,3,] *Hamid Sharifzadeh*[2]

[1] Leibniz-Centre General Linguistics (Leibniz-ZAS), Berlin
[2] Unitec Institute of Technology, Auckland, New Zealand
[3] Humboldt Universität zu Berlin, Berlin

sardhaei@leibniz-zas.de, zygis@leibniz-zas.de, hsharifzadeh@unitec.ac.nz

Speakers in their communications integrate auditory and visual information produced by face and body to communicate intended meanings. One of the core questions on the association between visual and spoken components of language is how various visual movements accompanying the speech contribute to the way the speech is uttered i.e., speech prosody. A substantial number of studies have demonstrated the prominent contribution of facial expressions and other forms of visual correlates to the auditory properties of prosody in voiced speech [1], [2], [3] [4], [5]. However, the literature has merely questioned what happens when F0, the most important prosodic cue, is absent from the acoustic signal, as is the case in whispered speech. A possible answer to this question can be interpreted in terms of trading relations. While on one hand, according to trade-off hypothesis, different modalities can compensate for the absence or reduced occurrence of another based on the requirements of situational constrains, on the other hand, hand-in-hand hypothesis views the relation between gestures and speech in parallel or redundant rather than compensatory in the sense that gestures basically express information that can be derived from the spoken content alone [6]. Building on these two hypotheses, we address the relation between orofacial expressions and acoustic cues of prosody under two specific communicative constraints. First, we examine if the relation between orofacial gestures and acoustic cues of intonation changes in two speech modes of normal and whispered speech differing in the presence or absence of fundamental frequency. Second, based on previous empirical studies reporting the effect of face masks on acoustic deterioration of an utterance [7], [8], [9], we investigate if various orofacial gestures such as the movements of eyebrows, lip aperture, and eye-squint are affected by wearing a protective face mask in comparison to a condition where the speakers do not use a facial mask. We will also study the extent the orofacial expressions interact with prosody in voiced vs. whispered to express intonational differences between questions with a rising intonation and statements with a falling F0.

we hypothesize that speakers may use more pronounced facial expressions in whispered speech than in voiced speech and in questions than statements. Assessing acoustic cues signaling intonation of sentence type, we predict longer duration of words, and reduced intensity when interlocutors whisper. Taken together, we hypothesize that speakers will intensify their oro-facial expressions when confronted with "marked" conditions, i.e., when they whisper, wear face masks, and produce questions. Finally, we will examine to what extent the gestural and acoustic parameters correlate. The analysis of gestures in parallel with acoustic signal will allow us to test if facial gestures compensate for the distorted acoustics signals including intensity and duration.

To this end, we ran an experiment with 10 Persian speakers producing 10 pairs of statements and questions in two modes of normal and whispered speech with and without face masks. Each sentence of the stimuli consisted of 4 words, with a bisyllabic target word in the final position. The second syllable of all the target words carrying the main stress, had a CVC structure started with a bilabial stop /p/, /b/, or /m/ and followed by the vowel /a/, /ɛ/ or /ɪ/. We tracked the movements of eyebrows, lips, as well as opening of eyes by a feature combination of OpenCV [10], a library of python binding, and the cross-platform Dlib [11] facial landmark detector to achieve an accurate facial landmark detection in the recorded videos.

Based on the results of linear mixed effect models, both left and right eyebrows are more raised in whispered than normal speech, in questions than statements, and when wearing a facial mask . The lip opening was found to be larger in the whispered mode of speech than normal speech ($t= 5.759$, $p<.001$) and larger in questions than in statements ($t= 3.555$, $p<.01$). For the lip aperture, the significant interaction between Sentence Type * Speech Mode ($t= -3.311$, $p<.001$) was reflected in a smaller difference between whispered and normal speech when statements are produced. The results also revealed a significant influence of speech mode on eye opening: eyes are opened larger in whispered than normal speech ($t= 3.075$, $p < 0.01$ for the left eye, and $t= 2.409$, $p < .05$ for the right eye). They were also larger when producing questions as opposed to statements ($t= 6.443$, $p < .0001$ for the left eye, and $t= 5.563$, $p < .0001$ for the right eye). From acoustic perspective, sentence-final words were longer in whispered than normal speech mode ($t= 3.684$, $p=001$).

Based on the outcomes of this study, depending on the communicative condition, the relationship between acoustic parameters and orofacial gestures can be in favor of one or the other hypothesis. More pronounced orofacial gestures and longer word duration may compensate for the lack of F0 supporting trade-off hypothesis. On the other hand, orofacial gestures can be also realized in parallel with longer acoustic duration which in turn supports hand-in-hand hypothesis.

Index Terms: oro-facial gesture, face mask, whispering

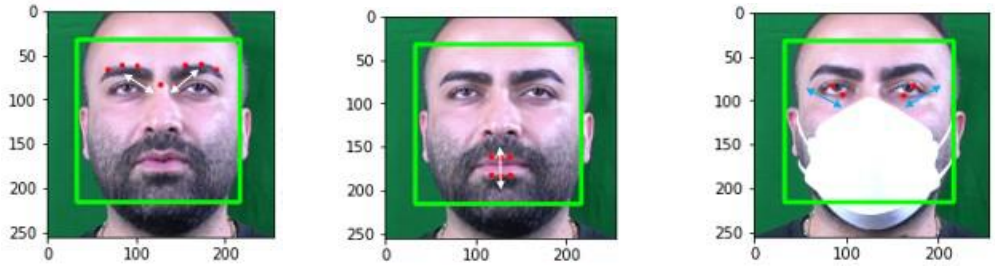
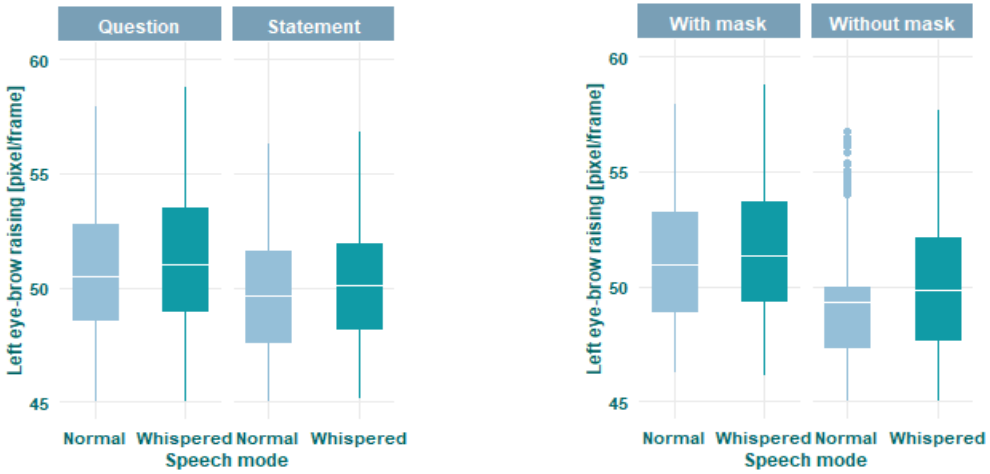
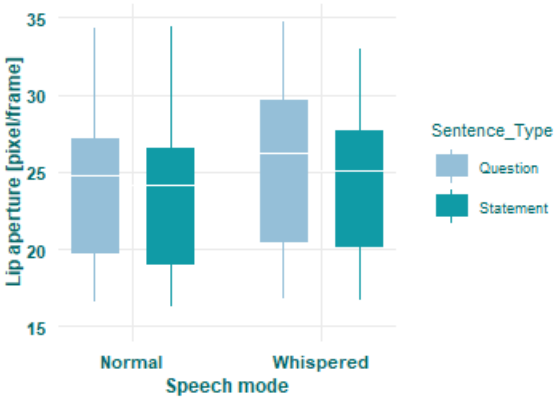


Figure 1: *Key points for the measurement of the eyebrow rising, lip aperture, and eye squint*



*Left eyebrow raising in the sentence-final word: interaction between **speech mode** and **sentence type** (left) as well as **speech mode** and **mask condition** (right).*



*Lip aperture in the in sentence-final word: interaction between **speech mode** and **sentence type***

References

[1] Guellaï, B., Langus, A., & Nespor, M. (2014). "Prosody in the hands of the speaker," Frontiers in Psychology 10, 1–8.

[2] Krahmer, E. J., and Swerts, M. (2007). "The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception," Journal of Memory and Language 57, 396–414.

[3] McNeill, D., Quek, F., McCullough, K. E., Duncan, S. D., Furuyama, N., Bryll, R., & Ansari, R. (2001). "Catchments, prosody and discourse," Gesture 1, 9-33.

[4] Mendoza-Denton, N., and Jannedy, S. (2011). "Semiotic layering through gesture and intonation: a case study of complementary and supplementary multimodality in political speech," Journal of English Linguistics 39, 265–299.

[5] Krahmer, E. J., and Swerts, M. (2009). "Audiovisual prosody—introduction to the special issue," Language and speech 52, 129-133.

[6] De Ruiter, J. P., Bangerter, A., & Dings, P (2021). "The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis," Topics in cognitive science 4, 232-248.

[7] Llamas, C., Harrison, P., Donnelly, D., & Watt, D. (2008). "Effects of different types of face coverings on speech acoustics and intelligibility," York Papers in Linguistics Series 2, 80–104.

[8] Maryn, Y., Wuyts, F., L., & Zarowski, A., J. (2021). "Are Acoustic Markers of Voice and Speech Signals Affected by Nose-and-Mouth-Covering Respiratory Protective Masks?," Journal of Voice.

[9] Porschmann, C., Lubeck, T., J., & Arend, M. (2020). "Impact of face masks on voice radiation," Journal of the Acoustical Society of America 148, 3663–3670.

[10] OpenCV (2015), Open-Source Computer Vision Library.

[11] King, D.E. (2009). "Dlib-ml: a machine learning toolkit," Journal of Machine learning research 10, 1755-1758.